Multivariate Statistics in Ecology and Quantitative Genetics
10. Quantitative Traits Loci (QTL) Mapping

Dirk Metzler & Martin Hutzenthaler

http://evol.bio.lmu.de/_statgen

1. June 2010

Contents

Introduction

Crossing Schemes

QTL model assumptions

Single-QTL analysis

LOD score

Interval mapping

More than one QTL

K.W. Broman, S. Sen (2009) *A guide to QTL Mapping with R/qtl*.
Springer, New York.

Contents

Introduction

- Crossing Schemes
- QTL model assumptions

Single-QTL analysis

- LOD score
- Interval mapping

More than one QTL

Contents

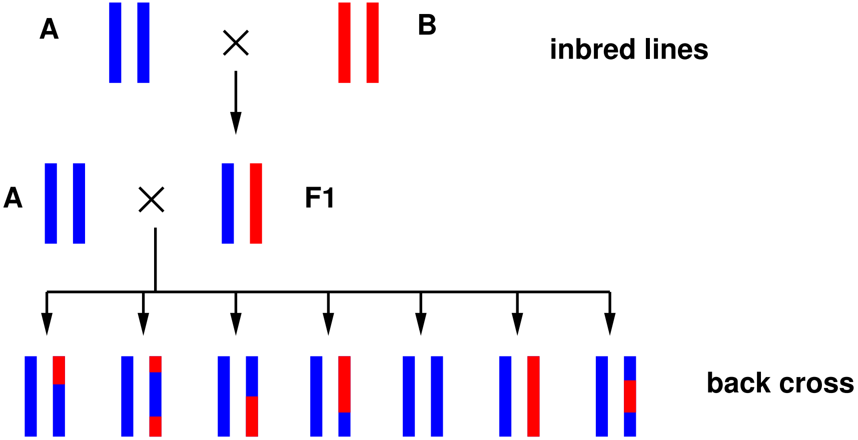
Introduction

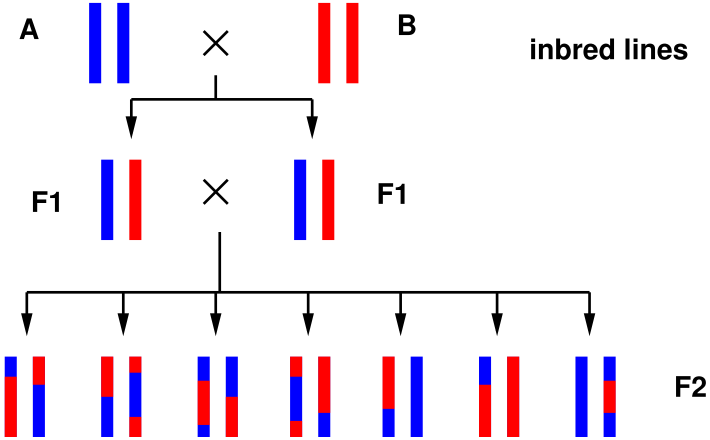
- Crossing Schemes
- QTL model assumptions

Single-QTL analysis

- LOD score
- Interval mapping

More than one QTL


A
×
B
inbred lines
A
×
F1
back cross



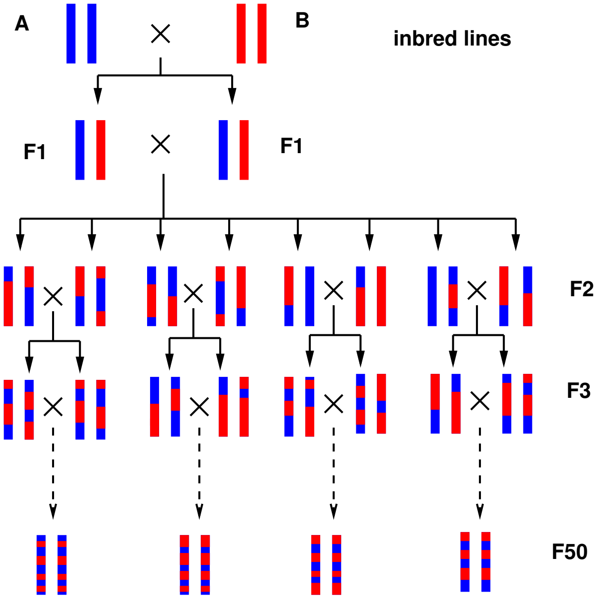
A × B

inbred lines

F1 × F1

F2

intercrosses



Recombinant Inbred Lines (RILs)

Contents

Assume that p sites have an influence on the quantitative trait y of interest and denote an individual's genotype at these sites by $g = (g_1, g_2, \ldots, g_p)$

$$\begin{aligned} \mu_g &:= \mathbb{E}(y|g) \\ \sigma_g^2 &:= \mathrm{var}(y|g) \\ \text{we assume: } y|g &\sim \mathcal{N}(\mu_g, \sigma_g^2) \\ \text{additive model: } \mu_g &= \mu + \sum_{j=1}^{p} z_j \cdot \Delta_j, \end{aligned}$$

whereas z_j is 0 or 1 according to the genotype of g_j, and Δ_j is the effect of the QTL at position j.

In a strict sense, *epistasis* means that the effect of a mutation can be masked by a mutation at a different loci.

However, in the context of QTL mapping, the word epistasis if often used to express that there is a non-additive interaction between two loci.

In a strict sense, *epistasis* means that the effect of a mutation can be masked by a mutation at a different loci.

However, in the context of QTL mapping, the word epistasis if often used to express that there is a non-additive interaction between two loci.

Main problem: We do not know where the QTLs are. We only have genetic markers to determine for several sites whether the stem from A or B.

Contents

Introduction

- Crossing Schemes
- QTL model assumptions

Single-QTL analysis

- LOD score
- Interval mapping

More than one QTL

Contents

Introduction

- Crossing Schemes
- QTL model assumptions

Single-QTL analysis

- LOD score
- Interval mapping

More than one QTL

Assume a backcross experiment with n F2 individuals
Let $y = (y_1, \ldots, y_n)$ be their phenotypes for the trait of interest.

Null hypothesis H_0: no QTL
Residual sum of squares under H_0:

$$\mathrm{RSS}_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Very simple alternative H_1: single QTL at marker position i

$$y_i | g_i \sim \mathcal{N}(\mu_{g_i}, \sigma^2)$$

Likelihood function:

$$\begin{aligned} L_1(\mu_{AA}, \mu_{AB}, \sigma^2) &= \Pr(y|\text{QTL marker}, \mu_{AA}, \mu_{AB}, \sigma^2) \\ &= \Pi_{i=1}^{n} \phi(y_i; \mu_{g_i}, \sigma^2), \end{aligned}$$

whereas ϕ is the density of the normal distribution.

The maximal likelihood under H_1 is $\mathrm{RSS}_1^{-n/2}$, with

$$\mathrm{RSS}_1 = \sum_{i=1}^{n} (y_i - \widehat{\mu_{g_i}})^2 \, .$$

The LOD score is the $\log_{10}$ of the likelihood ratio of H_1 and H_0:

$$\mathrm{LOD} = \frac{n}{2} \log_{10} \left(\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1} \right)$$

The LOD score is traditionally used in QTL mapping. However, it is equivalent to the classical anova F-statistic:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{df}}{\mathrm{RSS}_1/(n - \mathrm{df} - 1)} = \left(10^{2 \cdot \mathrm{LOD}/n} - 1\right) \cdot \frac{n - \mathrm{df} - 1}{\mathrm{df}}$$

$$\mathrm{LOD} = \frac{n}{2} \log_{10}\left(\frac{F \cdot \mathrm{df}}{n - \mathrm{df} + 1} + 1\right)$$

So, if the marker positions are our our candidates for the QTLs we just perform anovas.

Contents

Introduction

- Crossing Schemes
- QTL model assumptions

Single-QTL analysis

- LOD score
- Interval mapping

More than one QTL

- The QTLs may be between the marker positions, and their genotypes can only be estimated from the markers.

- The QTLs may be between the marker positions, and their genotypes can only be estimated from the markers.
- Let M_i be the multipoint marker genotype and g_i be the QTL genotype of individual i, and

  $$p_{ij} := \Pr(g_i = j | M_i).$$

  (Computation uses recombination rates.)

- The QTLs may be between the marker positions, and their genotypes can only be estimated from the markers.
- Let M_i be the multipoint marker genotype and g_i be the QTL genotype of individual i, and

$$p_{ij} := \Pr(g_i = j | M_i).$$

  (Computation uses recombination rates.)
- Probability density of an individual's phenotype is a mixture of normal distribution densities:

$$\sum_j p_{ij} \cdot \phi(y_i; \mu_j, \sigma^2)$$

EM algorithm for ML-estimation of μ_j and σ

Start with initial estimates $\mu_j^{(0)}$ and $\sigma^{(0)}$ and iterate the following steps for $s = 1, \ldots, N$:

E-step

$$\begin{aligned} w_{ij}^{(s)} &:= \Pr(g_i = j | M_i, y_i, \mu_j^{(s-1)}, \sigma^{(s-1)}) \\ &= \frac{p_{ij}\phi(y_i; \mu_j^{(s-1)}, \sigma^{(s-1)})}{\sum_k p_{ik}\phi(y_i; \mu_k^{(s-1)}, \sigma^{(s-1)})} \end{aligned}$$

M-step

$$\begin{aligned} \mu_j^{(s)} &:= \sum_i w_{ij}^{(s)} y_i / \sum_i w_{ij}^{(s)} \\ \sigma^{(s)} &:= \sqrt{\sum_{ij} w_{ij}^{(s)} (y_i - \mu_j^{(s)})^2 / n} \end{aligned}$$

The aim of the EM algorithm is that $\mu_j^{(s)}$ and $\sigma^{(s)}$ converge against the ML estimators $\widehat{\mu}$ and $\widehat{\sigma}$.

The aim of the EM algorithm is that $\mu_j^{(s)}$ and $\sigma^{(s)}$ converge against the ML estimators $\widehat{\mu}$ and $\widehat{\sigma}$.

Then, the LOD score can be computed:

$$\mathrm{LOD} = \log_{10}\left(\frac{\prod_i \sum_j p_{ij}\phi(y_i; \widehat{\mu}_j, \widehat{\sigma}^2}{\prod_i \phi(y_i; \widehat{\mu}_0, \widehat{\sigma}_0^2}\right)$$

Sometimes EM can be very slow.
Haley-Knott (HK) regression is a fast approximation:
For each point i on the grid calculate $p_{ij} = \Pr(g_i = j | M_i)$ and estimate μ_j and σ by fitting a linear model

$$y_i | M_i \sim \mathcal{N}\left(\sum_j p_{ij}\mu_j, \sigma^2\right)$$

Sometimes EM can be very slow.
Haley-Knott (HK) regression is a fast approximation:
For each point i on the grid calculate $p_{ij} = \Pr(g_i = j|M_i)$ and estimate μ_j and σ by fitting a linear model

$$y_i|M_i \sim \mathcal{N}\left(\sum_j p_{ij}\mu_j, \sigma^2\right)$$

Extended Haley-Knott (EHK) regression: Takes into account that p_{ij} and μ_j have an influence on the variance:

$$y_i|M_i \sim \mathcal{N}\left(\sum_j p_{ij}\mu_j\ ,\ \sum_j p_{ij}\mu_j^2 - \left(\sum_j p_{ij}\mu_j\right)^2 + \sigma^2\right)$$

Which LOD scores are significant?

Which LOD scores are significant?
Assess this by a permutation test: shuffle the phenotype column.

Contents

Introduction

- Crossing Schemes
- QTL model assumptions

Single-QTL analysis

- LOD score
- Interval mapping

More than one QTL

Composite Interval Mapping While searching for a QTL in one interval use other markers as proxies for nearby QTLs. Thus, markers are used as covariates. Specify maximal number of covariates and how far they should be away from the interval under examination.

two-QTL models search for interacting pairs of QTLs. Same methods like in 1-QTL model: EM, HK, EHK

multiple QTLs When candidate loci are found, fit regression models allowing for interactions and do variable selection.